

AN UNBIASED SAMPLING AND ESTIMATION PROCEDURE FOR  
CREEL CENSUSES OF FISHERMEN

BU-102-M

D. S. Robson

January 15, 1959

Introduction

The management of many sports fisheries is partly dependent upon information obtained through a creel census of the fishermen. Various types of data are collected in a creel census, including number of fish caught, amount of fishing time expended on the catch, proportion of marked or tagged fish in the catch, and morphometric data on the captured fish. The primary objective, however, is ordinarily considered to be the estimation of fishing mortality, or total number of fish removed by fishermen, since information on proportion of marked fish and on the morphometric characteristics of the fish population is obtainable by other, more efficient field methods. We shall, therefore, restrict our attention here to the problem of estimating total catch and total fishing effort.

A complete census of fishermen over a lake or stream is almost impossible to obtain due to practical limitations on the number of field personnel; consequently, creel censuses are usually designed as sample censuses. A ratio-type method of estimation frequently employed to estimate the total catch consists essentially of applying an estimated catch rate (number of fish caught per fishing hour) to the estimated total effort. This method of estimation exploits the facts that number of fish caught per fisherman, as measured by individual fisherman interviews, is positively correlated with the number of hours fished, and that the total number of hours fished by all fishermen can be estimated from easily obtained counts of the number of fishermen present at randomly chosen times during the day. The sampling design to be described here is constructed specifically to provide the data for an unbiased ratio-type estimate of total catch. Such a design will, of course, also provide the data for the unbiased estimation of total fishing effort and of the sampling error variance of the estimators.

Design of the Sample

The population of fishermen present on the given lake or stream through

the fishing season is considered to be stratified through space and time. The total area of the fishery is partitioned into  $N$  geographic segments in such a way that each segment supports approximately the same amount of fishing. A fishing season is stratified into weeks, and the weeks further stratified into three periods consisting of the five week-day period and each of the two weekend days. Fishing pressure varies systematically through the season and is much heavier on weekends than on weekdays; the stratification through time removes this systematic variation from the sampling error variance. Week-day holidays would also be separated into individual strata.

Within a given week-day period including, say,  $D$  days, the creel census is to be conducted on each of  $d$  randomly selected days. A crew of  $n$  enumerators conducts the census on  $n$  randomly selected area segments, each enumerator spending the entire day in his assigned area keeping a continuous record of the number of fishermen present and recording each man's catch from the area. The  $n$  sample segments are selected without replacement but independently on each of the  $d$  days.

Counts of the number of fishermen present in the entire fishery are to be made by an  $(n+1)$ st man on each of the  $d$  selected census days. His procedure is to traverse the fishery systematically but starting at a randomly chosen place and a randomly chosen time, counting all fishermen as he proceeds. A relatively rapid means of transportation is employed for the counting trip so that  $K$  complete circuits of the fishery could be made by continuous travel throughout the fishing day. For the sample,  $k$  of these  $K$  possible starting times are randomly selected without replacement. If fishing pressure is known to be unevenly distributed through the fishing day then stratification of this sample of starting times should be used; for example, the day might be partitioned into a morning stratum and an afternoon stratum, with a sample of  $k_1$  being chosen from the  $K_1$  possible starting times in the morning and  $k_2$  from the  $K_2$  possibilities in the afternoon.

#### Symbolic Description of the Population and Sample Data

The  $N$  area segments of the fishery are regarded as being numbered from 1 to  $N$  so that a random sample of  $n$  segments is obtained by randomly choosing a combination of  $n$  integers from the first  $N$  integers. Such a combination of  $n$

integers will be denoted generically by the symbol  $J_n$ ; there are then  $\binom{N}{n}$  different subsets  $J_n$  of the set  $(1,2,\dots,N)$ , each of which is equally likely to be chosen for the sample. Similarly,  $I_d$  will be used to denote a combination of  $d$  integers chosen from  $(1,2,\dots,D)$  and  $H_k$  denotes a subset of  $k$  integers from the set  $(1,2,\dots,K)$ . A set notation of this sort is required for a precise description of samples from a finite population.

The information being sought for a particular portion (stratum) of the fishing season includes the total number  $X$  of man hours fished and the total number  $Y$  of fish caught. In terms of the population structure defined above,

$$X = \sum_{i=1}^D \sum_{j=1}^N X_{ij}$$

where  $X_{ij}$  is the total number of man hours fished on the  $i$ 'th day at the  $j$ 'th segment and  $Y$  is similarly defined in terms of the corresponding  $Y_{ij}$ . The number of man hours  $X_{ij}$  may, in turn, be expressed as the integral of a counting function  $c_{ij}(t)$  over the entire day, where  $c_{ij}(t)$  is the number of fishermen present at time  $t$  in the  $j$ 'th segment on day  $i$ . As indicated in Figure 1,  $c_{ij}(t)$  is a step function with discontinuities at the points in time when fishermen enter or depart from the stratum. Time is most conveniently measured from an origin defined by the start of the fishing day, and the length of the fishing day is assumed to be an integer ( $=K$ ) multiple of the number of hours required, say  $\beta$  hours, for the completion of a counting trip; consequently,  $0 \leq t \leq K\beta$ .

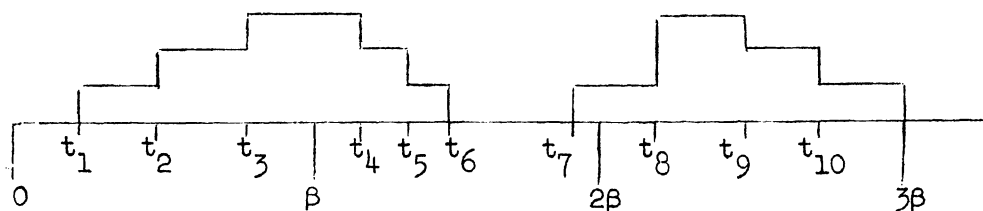


Figure 1. An example illustrating the form of the counting function  $c_{ij}(t)$

Since  $X_{ij}$  is the area under the graph of  $c_{ij}(t)$ ,

$$X_{ij} = \int_0^{K\beta} c_{ij}(t) dt.$$

If  $t_1, \dots, t_m$  are the points in time when fishermen enter or leave the segment, then

$$X_{ij} = \sum_{\alpha=0}^m (t_{\alpha+1} - t_{\alpha}) c_{ij}(t_{\alpha})$$

where  $t_0=0$ ,  $t_{m+1}=K\beta$ , and where  $c_{ij}(t_{\alpha})$  is the number of fishermen present during the interval from  $t_{\alpha}$  to  $t_{\alpha+1}$ .

The information assembled in the sample creel census conducted by the  $n$  enumerators includes, for each of  $d$  randomly chosen days, the effort  $X_{ij}$  and the catch  $Y_{ij}$  in  $n$  randomly chosen sample segments. This may be rephrased in terms of the set notation defined earlier to state that  $X_{ij}$  and  $Y_{ij}$  are observed for each  $i$  and  $j$  such that  $i$  belongs to  $I_d$  and  $j$  belongs to  $J_n^{(i)}$ .

Also included in the sample are the fisherman counts made on  $k$  counting trips during each of the  $d$  census days. The  $k$  starting times are selected at random and without replacement from the possible times  $t=0, t=\beta, t=2\beta, \dots, t=(K-1)\beta$ . A counting trip during the  $h$ 'th time interval on day  $i$  will reach the  $j$ 'th area segment at some time  $t$ ,  $(h-1)\beta \leq t \leq h\beta$ , resulting in an observed count  $c_{ijh}$ . The total number of fishermen observed on all  $k$  trips on day  $i$  may then be expressed as

$$\sum_{h \in H_k} \sum_{j=1}^N c_{ijh}$$

### Estimation of Total Fishing Effort

A counting trip starting at time  $(h-1)\beta$  on day  $i$  is begun at a randomly chosen place on the route; this ensures that the  $j$ 'th area segment is "equally likely" to be visited at any time during the interval from  $(h-1)\beta$  to  $h\beta$ . The time  $t$  at which the  $j$ 'th segment is counted is therefore a uniformly distributed chance variable on this interval, and the count  $c_{ijh}(t)$  is a chance variable having an expected value of

$$E_{ijh} \left\{ c_{ijh}(t) \right\} = \frac{1}{\beta} \int_{(h-1)\beta}^{h\beta} c_{ijh}(t) dt$$

for a fixed  $i$ ,  $j$  and  $h$ . This function is, except for the factor  $\frac{1}{\beta}$ , the area under the graph of  $c_{ijh}(t)$  between  $(h-1)\beta$  and  $h\beta$ ; hence,  $\beta c_{ijh}$  is an estimate of the total man hours fished, say  $X_{ijh}$ , in the  $j$ 'th segment during the time  $(h-1)\beta$  to  $h\beta$  on day  $i$ . The starting time is also chosen at random, however, and the expected value of the estimate  $\hat{X}_{ijh} = \beta c_{ijh}$  over all possible starting times is

$$\begin{aligned} E_{ij} \left\{ \hat{X}_{ijh} \right\} &= \frac{1}{K} \sum_{h=1}^K X_{ijh} \\ &= \frac{X_{ij}}{K} \end{aligned}$$

Consequently, an unbiased estimate of  $X_{ij}$  is

$$\begin{aligned} \hat{X}_{ij} &= \frac{K}{k} \sum_{h \in H_k(i)} \hat{X}_{ijh} \\ &= \frac{K\beta}{k} \sum_{H_k(i)} c_{ijh} \end{aligned}$$

and an unbiased estimate of the total effort  $X_i$  on the  $i$ 'th census day is

$$\hat{X}_i = \sum_{j \in J_n(i)} X_{ij} + \sum_{j \in J_{N-n}(i)} \hat{X}_{ij}$$

Finally, since the  $d$  census days were also randomly selected, an unbiased estimate of the total fishing effort  $X$  over all  $D$  days is

$$\begin{aligned} \hat{X} &= \frac{D}{d} \sum_{i \in I_d} \hat{X}_i \\ &= \frac{D}{d} \sum_{I_d} \left\{ \sum_{J_n(i)} X_{ij} + \frac{K\beta}{k} \sum_{H_k(i)} \sum_{J_{N-n}(i)} c_{ijh} \right\} \end{aligned}$$

A noteworthy feature of the mechanics of this estimation procedure is that the counting man, himself, need not record his count by area segment if the time at which he reaches each of the  $n$  sample segments being censused is recorded. The continuous records of the census enumerators will give the counts  $c_{ijh}$  at the recorded times of visit, and subtraction of these from the count man's total will give the desired sum of the  $c_{ijh}$  over the  $N-n$  segments numbered in  $J_{N-n}^{(i)}$ .

Sampling error in  $\hat{X}$  is seen to arise from several sources corresponding to the several stages of sampling. First, there is the error in  $\beta c_{ijh}$  as an estimate of  $X_{ijh}$ ; second, ignoring this first error, there would still be error in

$$\tilde{X}_{ij} = \frac{K}{k} \sum_{H(i)} X_{ijh}$$

as an estimate of  $X_{ij}$ ; and finally, ignoring these errors, there would still be error in

$$\tilde{X} = \frac{D}{d} \sum_{I_d} X_i$$

as an estimate of  $X$ . The presence of these three kinds of error in the error of the estimate  $\hat{X}-X$  may be illustrated algebraically by the identity

$$\begin{aligned} \hat{X}-X &= \frac{D}{d} \sum_{I_d} \sum_{J_{N-n}^{(i)}} \left\{ \frac{K}{k} \sum_{H(i)} (\beta c_{ijh} - X_{ijh}) + \left( \frac{K}{k} \sum_{H(i)} X_{ijh} - X_{ij} \right) \right\} + \left( \frac{D}{d} \sum_{I_d} X_i - X \right) \\ &= \frac{DK}{dk} \sum_{I_d} \sum_{J_{N-n}^{(i)}} \sum_{H(i)} (\hat{X}_{ijh} - X_{ijh}) + \frac{D}{d} \sum_{I_d} \sum_{J_{N-n}^{(i)}} (\tilde{X}_{ij} - X_{ij}) + (\tilde{X} - X) . \end{aligned}$$

The error variance of  $\hat{X}$ ,  $\text{Var}(\hat{X}) = E(\hat{X}-X)^2$ , is, likewise, made up of three components since the three error of estimate components defined above are uncorrelated, and each has mean 0. Thus,

$$E(\hat{X}-X)^2 = E \left\{ \frac{DK}{dk} \sum_{I_d} \sum_{J_{N-n}^{(i)}} \sum_{H_k^{(i)}} (\beta c_{ijh} - X_{ijh}) \right\}^2 + E \left\{ \frac{D}{d} \sum_{I_d} \sum_{J_{N-n}^{(i)}} (\tilde{X}_{ij} - X_{ij}) \right\}^2 + E(\tilde{X}-X)^2$$

$$= V_1 + V_2 + V_3 \quad .$$

The third component of variance, due to fishing pressure differences between days, is easily seen to have the familiar form

$$V_3 = E(\tilde{X}-X)^2$$

$$= \frac{D(D-d)}{d(D-1)} \sum_1^D \left\{ X_i^2 - \frac{1}{D} (\sum_1^D X_i)^2 \right\}$$

The second component  $V_2$ , due to variation among periods within days, takes a more complicated form

$$V_2 = E \left\{ \frac{D}{d} \sum_{I_d} \sum_{J_{N-n}^{(i)}} (\tilde{X}_{ij} - X_{ij}) \right\}^2$$

$$= \frac{D(N-n)}{d} \left( \frac{K(K-k)}{k} \right) \sum_{i=1}^D \left\{ \sum_{j=1}^N \sigma_{ij}^2 + \frac{N-n-1}{N-1} \sum_{j \neq j'}^N \sigma_{ijj'} \right\}$$

where  $\sigma_{ij}^2$  is the variance among periods in the  $i$ 'th day and  $j$ 'th segment

$$\sigma_{ij}^2 = \frac{1}{K-1} \sum_{h=1}^K (X_{ijh} - \frac{1}{K} \sum_{h=1}^K X_{ijh})^2$$

and  $\sigma_{ijj'}$  is the covariance between the number of hours fished per period in the two segments  $j$  and  $j'$  on day  $i$ ,

$$\sigma_{ijj'} = \frac{1}{K-1} \sum_{h=1}^K (X_{ijh} - \frac{1}{K} \sum_{h=1}^K X_{ijh}) (X_{ij'h} - \frac{1}{K} \sum_{h=1}^K X_{ij'h}) \quad .$$

The first component  $V_1$ , due to variation within a  $\beta$ -hour period, is

$$V_1 = E \left\{ \frac{DK}{dk} \sum_{I_d} \sum_{H_k^{(i)}} \sum_{J_{N-n}^{(i)}} (\hat{X}_{ijh} - X_{ijh}) \right\}^2$$

$$= \frac{DK(N-n)}{dkN} \sum_{i=1}^D \sum_{h=1}^K \left\{ \sum_{j=1}^N \sigma_{ijh}^2 + \frac{N-n-1}{N-1} \sum_{j \neq j'}^N \sigma_{ijj'h} \right\}$$

where  $\sigma_{ijh}^2$  is the variance, through time, of the number of fishermen present in segment  $j$  between the times  $(h-1)\beta$  and  $h\beta$  on day  $i$ ,

$$\sigma_{ijh}^2 = \beta \int_{(h-1)\beta}^{h\beta} c_{ij}^2(t) dt - X_{ijh}^2$$

Notice that

$$\sum_{h=1}^K \sum_{j=1}^N \sigma_{ijh}^2 = \beta \sum_{j=1}^N \int_0^{K\beta} c_{ij}^2(t) dt - \sum_{j=1}^N \sum_{h=1}^K X_{ijh}^2$$

The covariance  $\sigma_{ijj'h}$  between the fisherman counts in segments  $j$  and  $j'$  in the  $h$ 'th interval of day  $i$  is dependent upon the distance between the two segments in terms of travel time. If  $\beta_{jj'}$  is the amount of time required to travel from segment  $j$  to segment  $j'$  then

$$\sigma_{ijj'h} = \beta \int_{(h-1)\beta}^{(h-1)\beta + \beta_{jj'}} c_{ij}(t + \beta - \beta_{jj'}) c_{ij'}(t) dt$$

$$+ \beta \int_{(h-1)\beta + \beta_{jj'}}^{h\beta} c_{ij}(t - \beta_{jj'}) c_{ij'}(t) dt - X_{ijh} X_{ij'h}$$

The first integral accounts for the cases where the counting trip is started at a point on the route from segment  $j$  to segment  $j'$ , and the second integral covers the cases in which the trip is started between segment  $j'$  and segment  $j$ .

Estimation of  $\text{Var}(\hat{X})$  may also be carried out in essentially three steps corresponding to the three components of error variance. First, the statistic



$$s_X = \frac{D(D-d)}{d(d-1)} \left\{ \sum_{I_d} \hat{X}_i^2 - \frac{1}{k} \left( \sum_{I_d} \hat{X}_i \right)^2 \right\}$$

is an unbiased estimate of  $\text{Var}(\hat{X}) - \frac{d}{D}(V_1 + V_2)$ ; and second, the statistic

$$\begin{aligned} \hat{V}_2 = & \left( \frac{D^2}{d^2} \right) \left( \frac{N-n}{N(n-1)} \right) \left( \frac{K(K-k)}{k(K-1)} \right) \sum_{I_d} \left\{ \sum_{J_n^{(i)}} \left[ \sum_{h=1}^K x_{ijh}^2 - \frac{1}{K} \left( \sum_{h=1}^K x_{ijh} \right)^2 \right] \right. \\ & \left. + \frac{N-n-1}{n-1} \left[ \sum_{h=1}^K \left( \sum_{J_n^{(i)}} x_{ijh} \right)^2 - \sum_{J_n^{(i)}} \sum_{h=1}^K x_{ijh}^2 - \frac{1}{K} \left( \sum_{J_n^{(i)}} x_{ijh} \right)^2 + \frac{1}{K} \sum_{J_n^{(i)}} x_{ijh}^2 \right] \right\} \end{aligned}$$

is an unbiased estimate, term for term, of the component  $V_2$ . For computational purposes, this estimator may be collapsed to

$$\begin{aligned} \hat{V}_2 = & \frac{D^2 K(N-n)(K-k)}{d^2 k N(n-1)(K-1)} \sum_{I_d} \left\{ (N-n-1) \left[ \sum_{j=1}^K \left( \sum_{J_n^{(i)}} x_{ijh} \right)^2 - \frac{1}{K} \left( \sum_{J_n^{(i)}} x_{ijh} \right)^2 \right] \right. \\ & \left. - (N-2n) \left[ \sum_{J_n^{(i)}} \sum_{h=1}^K x_{ijh}^2 - \frac{1}{K} \sum_{J_n^{(i)}} x_{ijh}^2 \right] \right\} ; \end{aligned}$$

however, the correspondence between the terms of the estimate and the terms of the parameter  $V_2$  are lost by this reduction. The within period component of error variance  $V_1$  may be estimated in several ways without bias but with varying degrees of precision and corresponding varying degrees of computational difficulty. The preferred, more precise estimate is

$$\begin{aligned} \hat{V}_1 = & \frac{D^2 K(N-n)}{d^2 k n} \sum_{I_d} \left[ \beta \sum_{J_n^{(i)}} \int_0^{K\beta} c_{ij}^2(t) dt - \sum_{h=1}^K \sum_{J_n^{(i)}} x_{ijh}^2 \right. \\ & + \frac{N-n-1}{n-1} \sum_{h=1}^K \sum_{J_n^{(i)}} \left\{ \beta \int_{(h-1)\beta}^{(h-1)\beta + \beta_{jj'}} c_{ij'}(t) c_{ij}(t + \beta - \beta_{jj'}) dt \right. \\ & \left. \left. + \beta \int_{(h-1)\beta + \beta_{jj'}}^{h\beta} c_{ij'}(t) c_{ij}(t - \beta_{jj'}) dt - x_{ijh} x_{ij'h} \right\} \right] \end{aligned}$$

Since the functions  $c_{ij}(t)$  are step functions, all integrals in this estimate may be expressed (and computed) as summations. The covariance term in  $\hat{V}_1$  involves  $dn(n-1)$  integrals, each of which essentially requires the construction of two graphs, a graph of the function  $c_{ij}(t)c_{ij}(t+\beta-\beta_{jj})$  and a graph of  $c_{ij}(t)c_{ij}(t-\beta_{jj})$ . A graph of  $c_{ij}(t)c_{ij}(t+\beta-\beta_{jj})$  may be constructed by first superimposing the graph of  $c_{ij}(t)$  on the graph of  $c_{ij}(t)$ , with the graph of  $c_{ij}(t)$  translated  $\beta-\beta_{jj}$  units to the right as illustrated in Figure 2. The graph of the product is then easily obtained as shown at the bottom of Figure 2 and the area under the graph computed in a straightforward manner. A graph of  $c_{ij}(t)c_{ij}(t-\beta_{jj})$  is obtained in a similar manner, translating  $c_{ij}(t)$  to the left  $\beta_{jj}$  units.

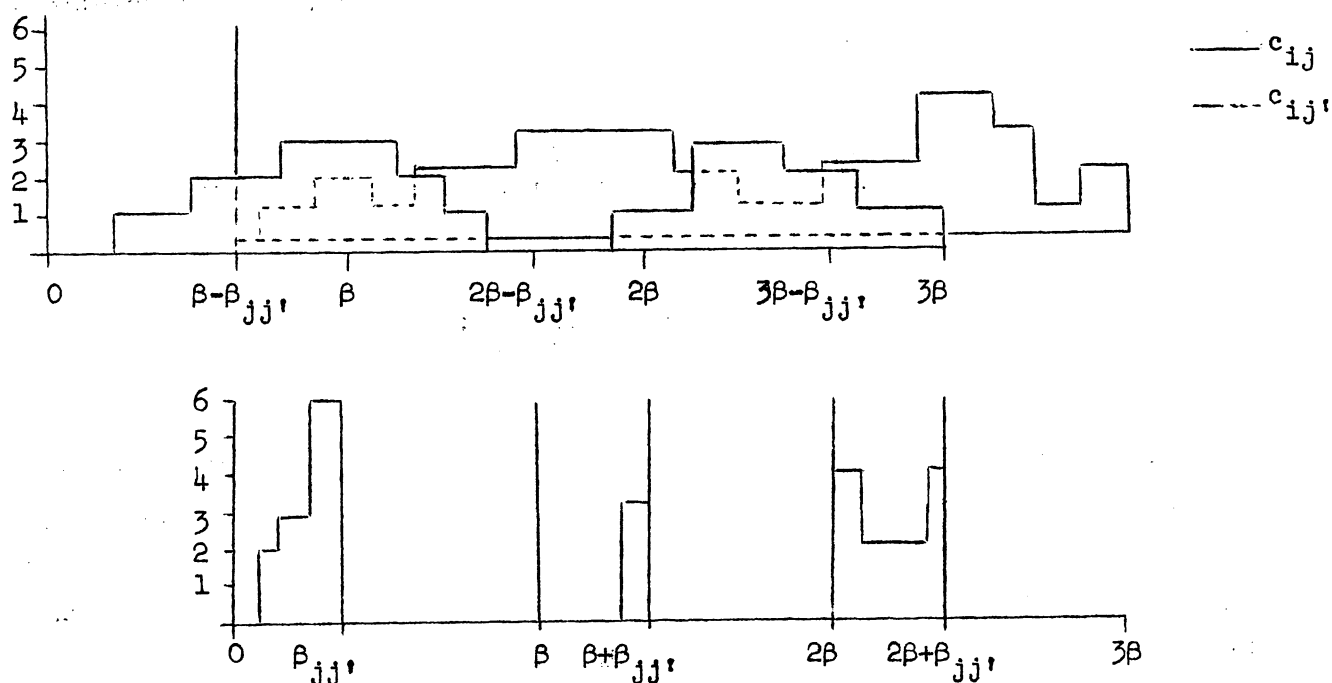


Figure 2. The top figure shows the graph of  $c_{ij}(t)$  superimposed on the graph of  $c_{ij}(t)$  with the latter translated  $\beta-\beta_{jj}$  units to the right. The lower figure plots the product of these two functions in their translated position, giving  $c_{ij}(t)c_{ij}(t+\beta-\beta_{jj})$  for the intervals  $(h-1)\beta$  to  $(h-1)\beta+\beta_{jj}$ .

If  $n$  is large then the number of such graphs, being of the order  $n^2$ , becomes prohibitive. An alternative estimator which utilizes the counts  $c_{ijh}$  made by the counting man to estimate the covariance term is

$$V_1 = \frac{D^2 K(N-n)}{d^2 k n} \sum_{I_d} \left[ \beta \sum_{J_n^{(i)}} \int_0^{K\beta} c_{ij}^2(t) dt - \sum_{h=1}^K \sum_{J_n^{(i)}} x_{ijh}^2 \right. \\ \left. + \frac{N-n-1}{n-1} \left\{ \frac{K\beta^2}{k} \sum_{H_k^{(i)}} \left[ \left( \sum_{J_n^{(i)}} c_{ijh} \right)^2 - \sum_{J_n^{(i)}} c_{ijh}^2 \right] - \sum_{h=1}^K \left[ \left( \sum_{J_n^{(i)}} x_{ijh} \right)^2 - \sum_{J_n^{(i)}} x_{ijh}^2 \right] \right\} \right].$$

This statistic is also an unbiased estimator of the within period component of the error variance, but is subject to a much larger sampling error than the estimator  $\hat{V}_1$ .

An unbiased estimator of the sampling variance of  $\hat{X}$  may now be formed by combining the three statistics  $S_X$ ,  $\hat{V}_2$  and  $\hat{V}_1$  into

$$\widehat{\text{Var}}(\hat{X}) = S_X + \frac{d}{D} (\hat{V}_1 + \hat{V}_2).$$

In addition, an estimator of the between-day component of the sampling variance is available now in the form of

$$\hat{V}_3 = S_X - \frac{D-d}{D} (\hat{V}_1 + \hat{V}_2).$$

The purpose of estimating sampling variance is, of course, to provide some measure of the precision of the estimates  $\hat{X}$ . Ordinarily, this measure is provided by the confidence interval

$$\hat{X} \pm 2 \sqrt{\widehat{\text{Var}}(\hat{X})}$$

based upon the assumption that  $\hat{X}$  is approximately normally distributed. Variance component estimates also permit the experimenter to estimate the change in precision he could expect by changing the sampling rate at any or all stages of sampling process.

# Estimation of the Total Catch

The purpose of constructing an unbiased estimator  $\hat{X}_1$  of the total effort  $X_1$  on the 1'th census day is primarily to provide the data for a ratio estimate of the total catch on day 1. The observed catch-rate in the sample of  $n$  area segments, multiplied by the estimate  $\hat{X}_1$ , forms an estimate of the total catch  $Y_1$ . This ratio estimate of  $Y_1$  is biased, however; and, as pointed out by Goodman and Hartley [1], the bias may be substantial. An alternative, unbiased ratio-type estimator has been constructed by Hartley and Ross [2] for the special case where the  $X_1$  is known; in the present notation, this Hartley-Ross estimator has the form

$$\tilde{Y}_1 = \left\{ \frac{1}{n} \sum_{J_n(i)} \frac{Y_{ij}}{X_{ij}} \right\} \left\{ X_1 - \frac{N-1}{n-1} \sum_{J_n(i)} X_{ij} \right\} + \frac{N-1}{n-1} \sum_{J_n(i)} Y_{ij} .$$

Since our estimator  $\hat{X}_1$  is unbiased and is, furthermore, statistically independent of the mean ratio

$$\begin{aligned} \bar{r}_1 &= \frac{1}{n} \sum_{J_n(i)} \frac{Y_{ij}}{X_{ij}} \\ &= \frac{1}{n} \sum_{J_n(i)} r_{ij} \end{aligned}$$

then the estimator

$$\hat{Y}_1 = \left\{ \frac{1}{n} \sum_{J_n(i)} r_{ij} \right\} \left\{ \hat{X}_1 - \frac{N-1}{n-1} \sum_{J_n(i)} X_{ij} \right\} + \frac{N-1}{n-1} \sum_{J_n(i)} Y_{ij}$$

is also unbiased. An unbiased estimate of the total catch  $Y$  for the entire  $D$  days is then

$$\hat{Y} = \frac{D}{d} \sum_{I_d} \hat{Y}_1 .$$

Again, the error of estimate may be expressed as a sum of uncorrelated error components with the result that sampling error variance is also made up of variance components. First, we observe that

$$\hat{Y}-Y = \frac{D}{d} \sum_{I_d} (\hat{Y}_1 - Y_1) + \left( \frac{D}{d} \sum_{I_d} Y_1 - Y \right)$$

or, letting

$$\tilde{Y} = \frac{D}{d} \sum_{I_d} Y_1 ,$$

then

$$\hat{Y}-Y = \frac{D}{d} \sum_{I_d} (\hat{Y}_1 - Y_1) + (\tilde{Y} - Y) .$$

Next, recalling the form of the Hartley-Ross estimator  $\tilde{Y}_i$ , we see that

$$\hat{Y}_i - Y_i = \frac{1}{n} \sum_{J_n^{(i)}} r_{ij} \sum_{j=1}^N (\hat{X}_{ij} - X_{ij}) + (\tilde{Y}_i - Y_i)$$

or, since  $\hat{X}_{ij} = X_{ij}$  for all  $j$  belonging to  $J_n^{(i)}$ ,

$$\hat{Y}_i - Y_i = \left( \frac{1}{n} \sum_{J_n^{(i)}} r_{ij} \right) \sum_{J_{N-n}^{(i)}} (\hat{X}_{ij} - X_{ij}) + (\tilde{Y}_i - Y_i) .$$

Finally, from the previous section, we have

$$\hat{X}_{ij} - X_{ij} = \frac{K}{k} \sum_{H_k^{(i)}} (\hat{X}_{ijh} - X_{ijh}) + (\tilde{X}_{ij} - X_{ij})$$

so that

$$\begin{aligned} \hat{Y}-Y &= \frac{DK}{dkn} \sum_{I_d} \left( \sum_{J_n^{(i)}} r_{ij} \right) \sum_{J_{N-n}^{(i)}} \sum_{H_k^{(i)}} (\hat{X}_{ijh} - X_{ijh}) \\ &\quad + \frac{D}{dn} \sum_{I_d} \left( \sum_{J_n^{(i)}} r_{ij} \right) \sum_{J_{N-n}^{(i)}} (\tilde{X}_{ij} - X_{ij}) \\ &\quad + \frac{D}{d} \sum_{I_d} (\tilde{Y}_1 - Y_1) + (\tilde{Y} - Y) . \end{aligned}$$

The four components of the error of estimate each have an expected value of 0 and, while they are not statistically independent, they are uncorrelated with one another. Consequently, the sampling variance of  $\hat{Y}$  is expressible as the sum of four components, say

$$\text{Var}(\hat{Y}) = C_1 + C_2 + C_3 + C_4 .$$

The last component  $C_4$  is simply

$$\begin{aligned} C_4 &= E(\tilde{Y} - Y)^2 \\ &= \frac{D(D-d)}{d(D-1)} \left\{ \sum_{i=1}^D Y_i^2 - \frac{1}{D} \left( \sum_{i=1}^D Y_i \right)^2 \right\} . \end{aligned}$$

and, as in the previous section, this component is ultimately estimated from the corresponding statistic

$$S_Y = \frac{D(D-d)}{d(d-1)} \left\{ \sum_{I_d} \hat{Y}_i^2 - \frac{1}{d} \left( \sum_{I_d} \hat{Y}_i \right)^2 \right\}$$

which is itself an estimate of

$$E(S_Y) = \text{Var}(\hat{Y}) - \frac{d}{D}(C_1 + C_2 + C_3) .$$

Thus, once the estimates  $\hat{C}_1$ ,  $\hat{C}_2$  and  $\hat{C}_3$  become available then  $\hat{C}_4$  may be estimated from  $S_Y$  by

$$\hat{C}_4 = S_Y \frac{D-d}{D} (\hat{C}_1 + \hat{C}_2 + \hat{C}_3) .$$

The third component is expressible as

$$\begin{aligned} C_3 &= E \left\{ \frac{D}{d} \sum_{I_d} (\tilde{Y}_i - Y_i)^2 \right\} \\ &= \frac{D}{d} \sum_{i=1}^D \text{Var}(\tilde{Y}_i) \end{aligned}$$

where  $\text{Var}(\tilde{Y}_i)$  is the variance of a Hartley-Ross ratio estimator for the  $i$ 'th day. Goodman and Hartley give this variance in its limiting form as  $N \rightarrow \infty$

$$\lim_{N \rightarrow \infty} \frac{\text{Var}(\tilde{Y}_i)}{N^2} = \frac{1}{n} \left\{ \text{Var}(Y_{ij}) + \bar{R}_i^2 \text{Var}(X_{ij}) - 2\bar{R}_i \text{Cov}(X_{ij}, Y_{ij}) \right. \\ \left. + \frac{1}{n-1} [\text{Var}(X_{ij})\text{Var}(r_{ij}) + \text{Cov}(X_{ij}, r_{ij})] \right\}$$

where  $\bar{R}_i$  is the mean value of the ratio  $r_{ij} = Y_{ij}/X_{ij}$  in the population. The exact variance of the Hartley-Ross estimator for the case  $N$  finite has been computed by Robson [3] and expressed in terms of multivariate, multipart cumulants which Tukey [4] calls polykays. Polykays are denoted by a set of vectors enclosed in square brackets, with the (integer) elements of the vectors representing the degree of the cumulant in each of the variables. In the present case the vectors contain three elements corresponding to the degree of each of the three variables  $X_{ij}$ ,  $Y_{ij}$  and  $r_{ij}$ . For example, the first cumulant of  $r_{ij}$  is expressed as  $\bar{R} = [(001)]'$ , the variance of  $r_{ij}$ ,

$$\text{Var}(r_{ij}) = \frac{1}{N-1} \sum_{j=1}^N (r_{ij} - \bar{R}_i)^2,$$

is denoted by  $[(002)]'$ , and the covariance between  $X_{ij}$  and  $r_{ij}$  is written  $[(101)]'$ . For the purpose at hand we shall merely use polykays as a convenient notation for reducing very tedious and uninteresting formulas into a compact form; the reader is referred to the cited papers by Tukey and Robson for the details of the algebra involved. The variance of  $\tilde{Y}_i$  is expressed in terms of polykays is

$$\text{Var}(\tilde{Y}) = \frac{N(N-n)}{n} \left\{ [(020)]' + [(001)(001)(200)]' - 2[(001)(110)]' \right. \\ \left. + \frac{N-1}{N(n-1)} [(200)(002)]' + \frac{N-n}{N(n-1)} [(101)(101)]' \right\}.$$

The subscript  $i$  has been omitted in this formula. Each term appearing in the formula represents a rather formidable polynomial function of the moments, as will be seen below in the estimation formulas; for interpretive purposes, however,

it may be noted that as N gets large this variance approaches term by term to the limiting value given earlier.

An unbiased estimate of  $\text{Var}(\tilde{Y}_1)$  is obtained by substituting into the above formula the unbiased estimates of the polykays. Computing formulas for these estimates are given below. All subscripts have been omitted, and sums are understood to extend over the set  $J_n^{(1)}$  of sample segment numbers; in addition, the product  $n(n-1)\cdots(n-S+1)$  is abbreviated to  $(n)_S$ .

$$\begin{aligned}
 [(020)] &= \frac{1}{(n)_2} \left\{ n\sum Y^2 - (\sum Y)^2 \right\} \\
 [(001)(001)(200)] &= \frac{1}{(n)_4} \left\{ (n-2)\sum X^2(\sum r)^2 - 2(n-1)\sum XY\sum r - (n-2)\sum X^2\sum r^2 + 2n\sum Y^2 \right. \\
 &\quad \left. - 4\sum X\sum Y\sum r - 2(\sum Y)^2 + (\sum X)^2\sum r^2 + 4\sum X\sum Y\sum r - (\sum X)^2(\sum r)^2 \right\} \\
 [(001)(110)] &= \frac{1}{(n)_3} \left\{ (n-1)\sum XY\sum r - n\sum Y^2 + \sum X\sum Y\sum r + (\sum Y)^2 - \sum X\sum Y\sum r \right\} \\
 [(200)(002)] &= \frac{1}{(n)_4} \left\{ (n^2-3n+1)\sum X^2\sum r^2 - n(n-1)\sum Y^2 - (n-2)(\sum X)^2\sum r^2 \right. \\
 &\quad \left. + 2(n-1)\sum X\sum Y\sum r - (n-2)\sum X^2(\sum r)^2 + 2(n-1)\sum XY\sum r \right. \\
 &\quad \left. + 2(\sum Y)^2 - 4\sum X\sum Y\sum r + (\sum X)^2(\sum r)^2 \right\} \\
 [(101)(101)] &= \frac{1}{(n)_4} \left\{ (n-1)(n-2)(\sum Y)^2 - n(n-1)\sum Y^2 + 2(n-1)(\sum X\sum Y\sum r \right. \\
 &\quad \left. + \sum XY\sum r - \sum X\sum Y\sum r) + \sum X^2\sum r^2 - \sum X^2(\sum r)^2 \right. \\
 &\quad \left. - (\sum X)^2\sum r^2 + (\sum X)^2(\sum r)^2 \right\}
 \end{aligned}$$

These same expressions with n replaced by N and with sums extending over the entire range, 1 to N, define the population polykays appearing in the formula for  $\text{Var}(\tilde{Y}_1)$ . With these computing formulas an unbiased estimate  $\widehat{\text{Var}}(\tilde{Y}_1)$  is obtained for each of the d sample days to give the following unbiased estimate of  $C_3$

$$\hat{C}_3 = \frac{D}{d} \sum_{I_d} \widehat{\text{Var}}(\tilde{Y}_1) .$$



The second component  $C_2$  of error variance is expressible in terms of the between-period variances  $\sigma_{ij}^2$  and covariances  $\sigma_{ijj'}$ , defined earlier for expressing the variance component  $V_2$ ; thus,

$$\begin{aligned}
 C_2 &= \frac{DK(K-k)(N-n)}{dknN(N-1)(N-2)} \sum_{i=1}^D \left\{ (n-1) \left( \sum_{j=1}^N r_{ij} \right) \left[ \left( \sum_{j=1}^N r_{ij} \right) \left( \sum_{j=1}^N \sigma_{ij}^2 \right) - 2 \sum_{j=1}^N r_{ij} \sigma_{ij}^2 \right] \right. \\
 &\quad + (N-n-1) \left( \sum_{j=1}^N r_{ij}^2 \right) \left( \sum_{j=1}^N \sigma_{ij}^2 \right) - (N-2n) \sum_{j=1}^N r_{ij}^2 \sigma_{ij}^2 \\
 &\quad + \frac{n-1}{N-3} \left( \sum_{j=1}^N r_{ij} \right) \left[ \left( \sum_{j=1}^N r_{ij} \right) \left( \sum_{j \neq j'} \sigma_{ijj'} \right) - \sum_{j \neq j'} r_{ij} \sigma_{ijj'} \right] \\
 &\quad \left. + \frac{(N-n-2)}{N-3} \left( \sum_{j=1}^N r_{ij}^2 \right) \left( \sum_{j \neq j'} \sigma_{ijj'} \right) - \frac{N-2n-1}{N-3} \sum_{j \neq j'} r_{ij}^2 \sigma_{ijj'} \right\} \\
 &= \frac{DK(K-k)(N-n)}{dknN(N-1)} \sum_{i=1}^D \left\{ \sum_{j \neq j'} r_{ij}^2 \sigma_{ij}^2 + \frac{n-1}{N-2} \sum_{j \neq j', j''} r_{ij} r_{ij'} \sigma_{ijj''} \right. \\
 &\quad + \frac{N-n-1}{N-2} \sum_{j \neq j', j''} r_{ij}^2 \sigma_{ijj''} \\
 &\quad \left. + \frac{(n-1)(N-n-1)}{(N-2)(N-3)} \sum_{j \neq j', j'', j'''} r_{ij} r_{ij'} r_{ij''} \sigma_{ijj'''} \right\} .
 \end{aligned}$$

An unbiased estimator of  $C_2$ , which is unbiased term by term according to the second formula for  $C_2$ , is

$$\begin{aligned}
 \hat{C}_2 &= \frac{D^2 K(K-k)(N-n)}{d^2 kn^2} \sum_{I_d} \left\{ \frac{1}{n-1} [(\sum r_{ij}^2)(\sum \sigma_{ij}^2) - \sum r_{ij}^2 \sigma_{ij}^2] + \frac{1}{n-2} [(\sum r_{ij})^2 \sum \sigma_{ij}^2 \right. \\
 &\quad \left. - (\sum r_{ij}^2)(\sum \sigma_{ij}^2) - 2(\sum r_{ij})(\sum r_{ij} \sigma_{ij}^2) + 2 \sum r_{ij}^2 \sigma_{ij}^2] \right\}
 \end{aligned}$$

$$\begin{aligned}
 & + \frac{N-n-1}{(n-1)(n-2)} [ (\Sigma r_{ij}^2) (\Sigma_{j \neq j'} \sigma_{ijj'}) - \Sigma_{j \neq j'} r_{ij}^2 \sigma_{ijj'} ] \\
 & + \frac{N-n-1}{(n-2)(n-3)} [ (\Sigma r_{ij})^2 (\Sigma_{j \neq j'} \sigma_{ijj'}) - (\Sigma r_{ij}^2) (\Sigma_{j \neq j'} \sigma_{ijj'}) \\
 & - 2(\Sigma r_{ij}) (\Sigma_{j \neq j'} r_{ij} \sigma_{ijj'}) + 2 \Sigma_{j \neq j'} r_{ij}^2 \sigma_{ijj'} ] \}
 \end{aligned}$$

where all sums extend over the set  $J_n^{(i)}$  for day  $i$ . Computing formulas were given earlier for the terms  $\Sigma \sigma_{ij}^2$  and  $\Sigma \sigma_{ijj'}$ ; the only additional formulas now required are

$$\Sigma r_{ij}^2 \sigma_{ij}^2 = \frac{1}{K-1} [ \Sigma r_{ij}^2 \Sigma_{h=1}^K X_{ijh}^2 - \frac{1}{K} \Sigma Y_{ij}^2 ]$$

$$\Sigma r_{ij} \sigma_{ij}^2 = \frac{1}{K-1} [ \Sigma r_{ij} \Sigma_{h=1}^K X_{ijh}^2 - \frac{1}{K} \Sigma X_{ij} Y_{ij} ]$$

$$\begin{aligned}
 \Sigma r_{ij}^2 \sigma_{ijj'} &= \frac{1}{K-1} [ \Sigma_{h=1}^K \left\{ (\Sigma r_{ij}^2 X_{ijh}) (\Sigma X_{ijh}) - \Sigma r_{ij}^2 X_{ijh}^2 \right\} \\
 &\quad - \frac{1}{K} \left\{ (\Sigma r_{ij} Y_{ij}) (\Sigma X_{ij}) - \Sigma Y_{ij}^2 \right\} ]
 \end{aligned}$$

$$\begin{aligned}
 \Sigma r_{ij} \sigma_{ijj'} &= \frac{1}{K-1} [ \Sigma_{h=1}^K \left\{ (\Sigma r_{ij} X_{ijh}) (\Sigma X_{ijh}) - \Sigma r_{ij} X_{ijh}^2 \right\} \\
 &\quad - \frac{1}{K} \left\{ \Sigma X_{ij} \Sigma Y_{ij} - \Sigma X_{ij} Y_{ij} \right\} ]
 \end{aligned}$$

where, again, all sums extend over  $J_n^{(i)}$  unless otherwise indicated.

The first component of error variance,

$$C_1 = E \left\{ \frac{DK}{dkn} \sum_{I_d} \left( \sum_{J_n^{(i)}} r_{ij} \right) \sum_{J_{N-n}^{(i)}} \sum_{H_k^{(i)}} (\hat{X}_{ijh} - X_{ijh}) \right\}^2,$$

is expressible in terms of the variances  $\sigma_{ijh}^2$  and covariances  $\sigma_{ijj'h}$  within periods as

$$C_1 = \frac{DK(N-n)}{dknN(N-1)} \sum_{i=1}^D \sum_{h=1}^K \left\{ \sum_{j \neq j'}^N r_{ij}^2 \sigma_{ij'h}^2 + \frac{n-1}{N-2} \sum_{j \neq j', j''} r_{ij} r_{ij'} \sigma_{ij'jh}^2 \right. \\ \left. + \frac{N-n-1}{N-2} \left[ \sum_{j \neq j', j''} r_{ij}^2 \sigma_{ij'jh} \right. \right. \\ \left. \left. + \frac{n-1}{N-3} \sum_{j \neq j', j'', j'''} r_{ij} r_{ij'} \sigma_{ij'j''jh} \right] \right\}$$

A term by term unbiased estimator of  $C_1$  is then

$$\hat{C}_1 = \frac{D^2 K(N-n)}{d^2 kn^2} \sum_{I_d} \sum_{h=1}^K \left\{ \frac{1}{n-1} [(\sum r_{ij}^2)(\sum \sigma_{ijh}^2) - \sum r_{ij}^2 \sigma_{ijh}^2] \right. \\ \left. + \frac{1}{n-2} [(\sum r_{ij})^2 (\sum \sigma_{ijh}^2) - (\sum r_{ij}^2)(\sum \sigma_{ijh}^2 - 2(\sum r_{ij})(\sum r_{ij} \sigma_{ijh}^2) \right. \\ \left. + 2\sum r_{ij}^2 \sigma_{ijh}^2)] + \frac{N-n-1}{(n-1)(n-2)} [(\sum_{j'} r_{ij})(\sum_{j \neq j'} \sigma_{ijj'h}) - \sum_{j \neq j'} r_{ij}^2 \sigma_{ijj'h}] \right. \\ \left. + \frac{N-n-1}{(n-2)(n-3)} [(\sum r_{ij})^2 (\sum_{j \neq j'} \sigma_{ijj'h}) - (\sum r_{ij}^2)(\sum_{j \neq j'} \sigma_{ijj'h}) \right. \\ \left. - 2(\sum r_{ij})(\sum_{j \neq j'} r_{ij} \sigma_{ijj'h}) + 2 \sum_{j \neq j'} r_{ij}^2 \sigma_{ijj'h} \right] \left. \right\}$$

where all sums extend over the set  $J_n^{(i)}$  for day  $i$ . Notice, for computing purposes, that the sum over the index  $h$  may be performed first; for example,

$$\begin{aligned} \sum_{h=1}^K \sum_n^{(i)} r_{ij}^2 \sigma_{ijh}^2 &= \sum_n^{(i)} (r_{ij}^2) \left( \sum_{h=1}^K \sigma_{ijh}^2 \right) \\ &= \sum_n^{(i)} r_{ij}^2 \left[ \beta \int_0^{K\beta} c_{ij}^2(t) dt - \sum_{h=1}^K x_{ijh}^2 \right] . \end{aligned}$$

The availability of estimators of the four components of error variance now permits the estimation of  $\text{Var}(\hat{Y})$  in the form of  $\widehat{\text{Var}}(\hat{Y}) = \hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_4$ , and the effect on sampling error variance resulting from varying the sampling rates may now be estimated.

### Discussion

This sampling and estimation procedure has several advantages over creel census methods which have been used in the past; it also has several disadvantages, the least of which is complexity of the computational procedure. The unbiased property is the main advantage of this scheme; it is unbiased in the statistical sense that, given the data called for in the sampling model, the error of the estimate has a theoretical expectation equal to 0, and it is unbiased in the sense of interviewer and respondent bias since none of the information called for is dependent upon the memory or opinions of either party. The chief disadvantage of this scheme is that it represents inefficient use of time of the field personnel; an enumerator conducting the census in an area segment will be standing idle for a large part of the day if the turnover of fishermen is low. Clearly, a sampling scheme which permitted the enumerator to rove the fishery would result in greater fisherman contact per enumerator, and if this technique could be incorporated into a sampling and estimation procedure having the unbiased property mentioned above then it would be much preferred to the present scheme.

A second advantage of this procedure, though stemming from the unbiased property, is the availability of estimates of error variance and its components. Fishery biologists in general recognize the importance of attaching a measure of accuracy to point estimates, and strive always to devise sampling procedures which allow estimation of error variance. A sampling procedure which does not permit unbiased point estimation can only produce a biased estimate of the measure of accuracy. A disadvantage of the present scheme in this respect is that

estimability of error variance requires minimum sample sizes of  $d \geq 2$ ,  $k \geq 2$  and  $n \geq 4$ ; that is, 4 or more census enumerators are required to work 2 or more of the  $D$  days, and the counting man is required to make 2 or more trips on each of these 2 or more days. If the fishery is large enough or if several fisheries in an area are being surveyed then the use of 5 or more field men may be justified; their effort, in this case, could be distributed so as to keep them occupied during the entire  $D$ -day period with their days randomly allotted among the several fisheries or among strata of the large fishery.

#### References

- [1] Goodman, L. A., and Hartley, H. O., "The precision of unbiased ratio-type estimators," Journal of the American Statistical Association, 53 (1958), 491-508.
- [2] Hartley, H. O., and Ross, A., "Unbiased ratio estimators," Nature, 174 (1954), 270.
- [3] Robson, D. S., "Applications of multivariate polykays to the theory of unbiased ratio-type estimation," Journal of the American Statistical Association, 52 (1957), 511-522.
- [4] Tukey, J. W., "Keeping moment-like sampling computations simple," Annals of Mathematical Statistics, 27 (1956), 37-54.